ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language

Year: 2020 | Citations: 355 | Authors: Oyvind Tafjord, Bhavana Dalvi, Peter Clark

---

Abstract

Transformers have been shown to emulate logical deduction over natural language theories (logical rules expressed in natural language), reliably assigning true/false labels to candidate implications. However, their ability to generate implications of a theory has not yet been demonstrated, and methods for reconstructing proofs of answers are imperfect. In this work we show that a generative model, called ProofWriter, can reliably generate both implications of a theory and the natural language proof(s) that support them. In particular, iterating a 1-step implication generator results in proofs that are highly reliable, and represent actual model decisions (rather than post-hoc rationalizations). On the RuleTaker dataset, the accuracy of ProofWriter's proofs exceed previous methods by +9% absolute, and in a way that generalizes to proof depths unseen in training and on out-of-domain problems. We also show that generative techniques can perform a type of abduction with high precision: Given a theory and an unprovable conclusion, identify a missing fact that allows the conclusion to be proved, along with a proof. These results significantly improve the viability of neural methods for systematically reasoning over natural language.